Online Sales Data Clustering & Association Pattern Mining

Arjun Chowhan
Artificial Intelligence & Machine Learning
Lambton College
Toronto, Canada
C0850490@mylambton.ca

Niteesha Balla
Artificial Intelligence & Machine Learning
Lambton College
Toronto, Canada
C0850488@mylambton.ca

Dimble Scaria
Artificial Intelligence & Machine Learning
Toronto, Canada
C0838227@mylambton.ca

Mohammed Omer Shaik
Artificial Intelligence & Machine
Lambton College
Toronto, Canada
C0828780@mylambton.ca

***Abstract*— The project aims on finding interesting patterns in online sales dataset and explore on association between different products using association pattern mining to make useful suggestions for the business owners and hence help the retail industry.**

Keywords—clustering, association pattern mining, exploratory data analysis.

I. Introduction

The gifts industry is an ever growing and flourishing sector of the market. People love sharing presents on their different occasions and it never ends. So, exploring more into this market and study the varying trend is very interesting and worth investing for business owners, manufactures and online retailers. Through our project we intend to find out interesting patterns in the sales data of an online retail company and find the association between sales of different products.

II. Motivation

We are intending to look up to a UK-based registered online retail company which mainly sells unique all-occasion gifts to wholesalers and the transactional data for over a period of 1 year, from 01/12/2010 to 09/12/2011 is taken into consideration for our analysis. The findings made through our analysis would be useful for the business owners and the retailors to design new strategies including providing offers for different target audience at different periods of time and investing at different markets according to the pattern changes.

III. Data Overview

Our dataset contains the details of sales of a UK based online wholesale company over a period of one year. It included 541909 records in total with 8 attributes namely:

- InvoiceNo
- StockCode
- Description
- Quantity
- InvoiceDate
- UnitPrice
- CustomerID
- Country

IV. Exploratory Data Analysis

According to IBM, “Exploratory Data Analysis is used to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.” [1].

This is a very important step in data mining through which the analyst gets to know more about data by the initial overview understanding the attributes and their datatypes, detecting the presence of errors, noise, and outlier in the data.

From the exploratory data analysis, presence of missing values and outliers were found in the dataset which was removed. This reduced the dataset into over 4,50,000 records which was considered for the further analysis. The box plot (fig 1) below represents the outliers and the variance in the dataset and fig 2 is the representation after removing outliers.

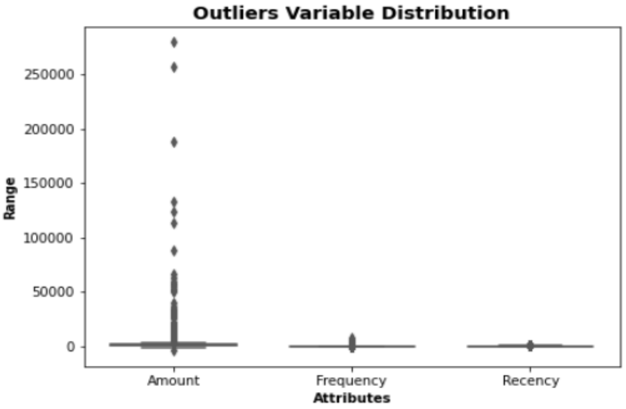


Fig 1. Box plot representing the outliers and variance in data

After the removal of the outlier the range of amount of amount reduced from 250000 to 12500. This helped in removing the high possible bias of the data so that we can get a more generalized result.

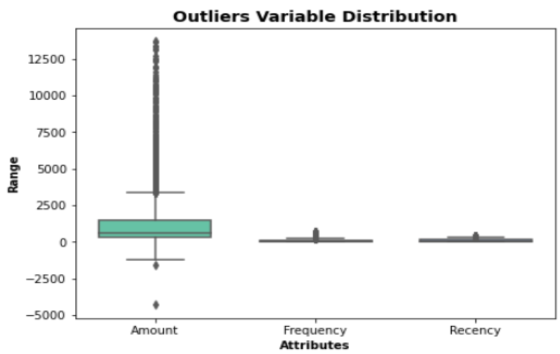


Fig 2. Box plot after removing the outliers

V. Data Visualization

Data visualization helps in better understanding a data. Rather than columns of number, visually appealing plots can better comprehend with the analyst helping him in making useful finds without much overhead and deriving better conclusions even faster.

A. Plotting the sales in different countries

The online retail company is focusing mainly on our European markets and Australia. The society there is more diverse due to heavy migrations making it a hot market for the gift industries through out the year. Among them the markets of countries namely, United Kingdom, France, Australia, Netherlands is found to perform way better than the other counties of the same region. The presence world famous happening, top metropolitan cities such as London, Melbourn, Paris can be a possible reason for this huge difference. So, focusing more on these markets and identifying its markets trends and investing more in this region would be a great idea. Also, expansion plans should be carefully planned for other countries in the list. Fig 3 gives a clear insight into this information.

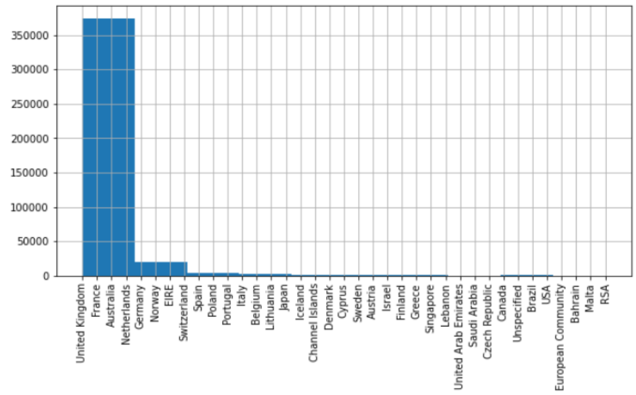


Fig. 3 Plotting showing sales in different counties.

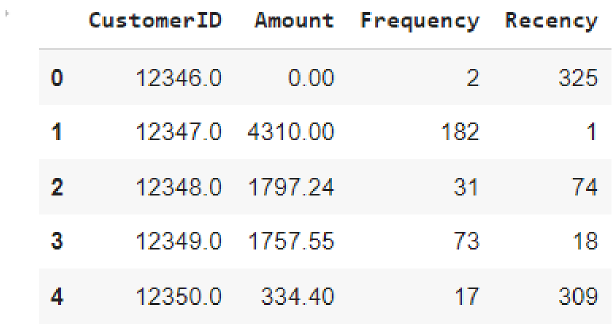
VI. Data PreProcessing

Data preprocessing is an important part of data mining. The data obtained is processed and converted into a format useful to achieve our aim at this step. The clustering modeling mainly focuses on 4 factors:

- R (Recency): Number of days since last purchase
- F (Frequency): Number of transactions
- M (Monetary): Total amount of transactions (revenue generated)

Through recency, we aim to analyse the durations between subsequent purchases. In frequency, we expect to analyse the frequency of purchase of each customer. Through this we could suggest possible business offers that can be given to customers to increase their frequency of purchase there by increasing the profit of the company. In the monetary aspect, we analysis we analysed how much each customer spend on gift purchase for the past one year.

Form the available data we had, we reduced the dataset into a dataset with four columns for or analysis which included customer_ID, Recency, Frequency and Monetary. The fig 4 shows a part of the reduced dataset.

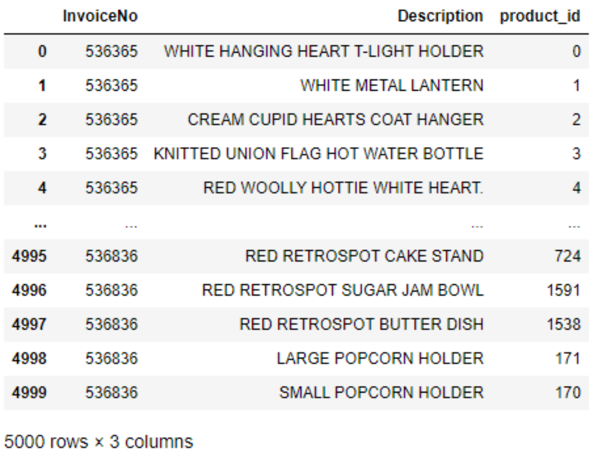


	CustomerID	Amount	Frequency	Recency
0	12346.0	0.00	2	325
1	12347.0	4310.00	182	1
2	12348.0	1797.24	31	74
3	12349.0	1757.55	73	18
4	12350.0	334.40	17	309

Fig. 4. Image of the reduced dataset

In association pattern mining we had to prepare the purchase list of each customer at each shopping from the existing dataset. The products were recorded as product descriptions in the original dataset. This had to be mapped as product codes so, for the ease of processing each of the products were given distinct product_ID which made our process so easy.

There were 4224 distinct products in our dataset and each of them were identified by a distinct code from 0 to 4224. Hence for association pattern mining the dataset was reduced to a dataframe (Fig. 5) containing customer ID product description and the product_ID.

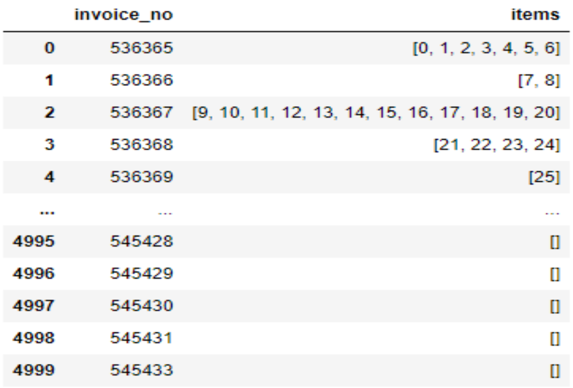


	InvoiceNo	Description	product_id
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	0
1	536365	WHITE METAL LANTERN	1
2	536365	CREAM CUPID HEARTS COAT HANGER	2
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	3
4	536365	RED WOOLLY HOTTIE WHITE HEART.	4
...	...	...	...
4995	536836	RED RETROSPOT CAKE STAND	724
4996	536836	RED RETROSPOT SUGAR JAM BOWL	1591
4997	536836	RED RETROSPOT BUTTER DISH	1538
4998	536836	LARGE POPCORN HOLDER	171
4999	536836	SMALL POPCORN HOLDER	170

5000 rows × 3 columns

Fig. 5 products identified with distinct product_ID

The purchase list of each customer on each deal is them made from this dataset. Because of our computational limitation our further research was done on a subset of the original dataset. The following figure (Fig.6) depicts the dataframe with items recorded with distinct product_IDs.

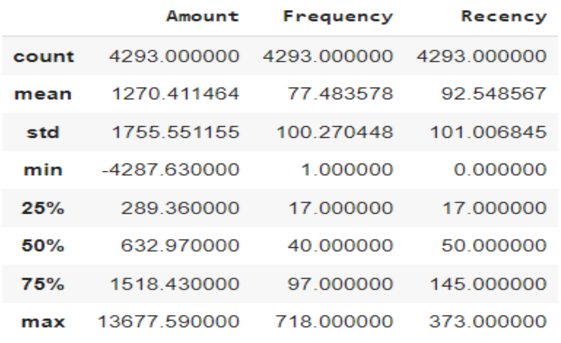


	invoice_no	items
0	536365	[0, 1, 2, 3, 4, 5, 6]
1	536366	[7, 8]
2	536367	[9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]
3	536368	[21, 22, 23, 24]
4	536369	[25]
...	...	...
4995	545428	[]
4996	545429	[]
4997	545430	[]
4998	545431	[]
4999	545433	[]

Fig 6. purchase list of each sale

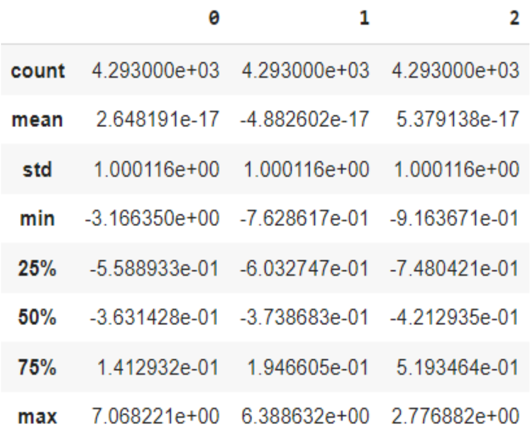
VII. FEATURE SCALING

If the datapoints spreads across a huge range, the feature scaling is done to normalize values. Here, we have used standardized scaling. This helps in bringing the datapoints closer making the identification of the clusters easier. This really helped in reducing the variance in the Monetary attribute in the dataset which we used in determining the contribution of each customer and hence bringing the datapoints together. The description of the dataset before feature scaling is described in first figure (fig 7) and after in second figure (fig 8).



	Amount	Frequency	Recency
count	4293.000000	4293.000000	4293.000000
mean	1270.411464	77.483578	92.548567
std	1755.551155	100.270448	101.006845
min	-4287.630000	1.000000	0.000000
25%	289.360000	17.000000	17.000000
50%	632.970000	40.000000	50.000000
75%	1518.430000	97.000000	145.000000
max	13677.590000	718.000000	373.000000

Fig 7. Description of the dataset before feature scaling



	0	1	2
count	4.293000e+03	4.293000e+03	4.293000e+03
mean	2.648191e-17	-4.882602e-17	5.379138e-17
std	1.000116e+00	1.000116e+00	1.000116e+00
min	-3.166350e+00	-7.628617e-01	-9.163671e-01
25%	-5.588933e-01	-6.032747e-01	-7.480421e-01
50%	-3.631428e-01	-3.738683e-01	-4.212935e-01
75%	1.412932e-01	1.946605e-01	5.193464e-01
max	7.068221e+00	6.388632e+00	2.776882e+00

Fig 8. Description of the dataset before feature scaling

VIII. MODELS

VIII.1 Clustering modelling

Analysing the data to find common patterns and trends among groups is very beneficial in business or sales analysis. Here we are using 2 different techniques of clustering to accomplish the task.

VIII.1.1 K Means Clustering

K Means Clustering is a simple clustering technique used in unsupervised machine learning for finding interesting patterns shown by the data. On applying K Means clustering, we successfully identified 3 clustering. The results were tried different number of clustering using the Elbow plot (Fig 9) and clustering the data into 3 clusters is found to be the optimal. Silhouette scoring was also done for different clusters and number of clusters of 3 is found to perform better with a Silhouette score of 0.50848.

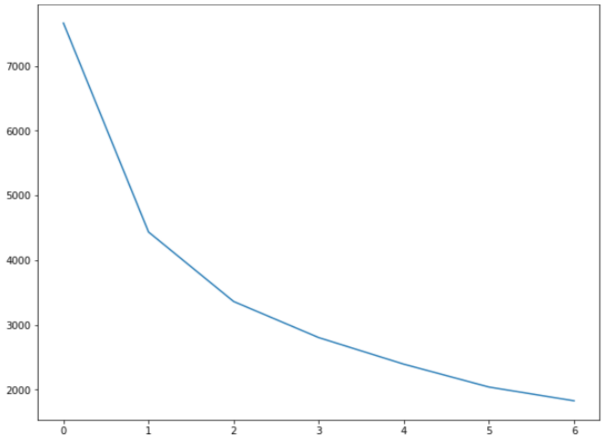


Fig 9. Elbow plot

VIII.1.1 Hierarchical Clustering

In hierarchical clustering, the number of clusters identified was also 3 just as in the K Means Clustering. The fig 10 shows the complete linkage hierarchical clustering plot. This result underlines the result obtained in K Means clustering.

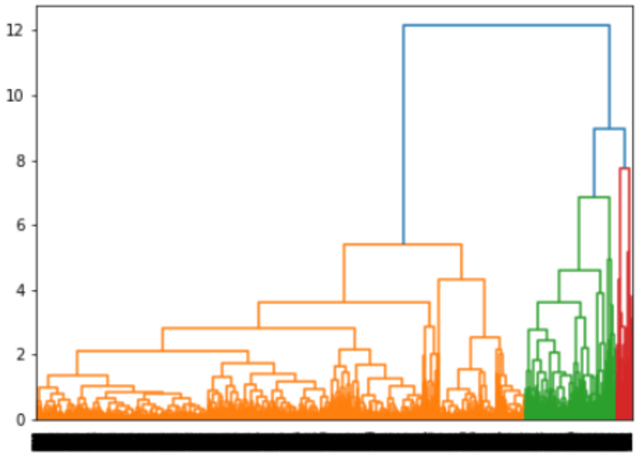


Fig 10. Complete Linkage hierarchical clustering plot

VIII.2 Association Pattern Modelling

The association pattern mining was a challenging part of the project. The huge size of the data was the main challenge we faced. So, as a part of initial analysis we worked on a subset of out original dataset of 5000 records. The model successfully run end to end but were not able to come up with good associations. This was due to the huge number of product varieties and the limited number of data records considered.

There were 4224 different products and comparing these 5000 records would show heavy variance in the data making it impossible to make association among products. We used Apriori algorithm for our pattern mining.

To come up with meaning full associations, we increased our data limit into 50,000 records and again carried out the process. And we successfully come up with some associations. On giving the minimum support as 0.02 to derive around 280 associations.

IX. INFERENCES

Some people are more into celebrations, gifts, and shopping but it is not same always. Also, the difference in culture, climate and lifestyle of the people do influence the gift industry.

Through our analysis, we were able to successfully identify 3 groups of customers form our online retails sales data. The first cluster is found to spend more on gifts and is contributing more to the sales. The same cluster is found to be more frequent into the gift stores. Fig 11 shows the difference in expenditure of the 3 groups. Frequency of shopping by different clusters is represented in fig 12.

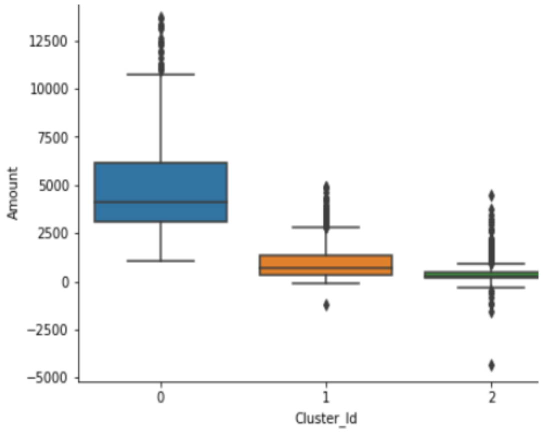


Fig 11. Box plot representing Amount spend by different clusters.

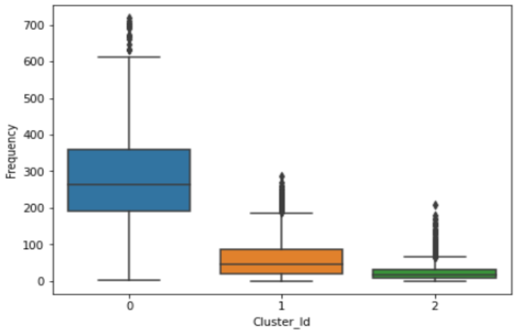


Fig 12. Box plot representing frequency in purchase by different customer clusters

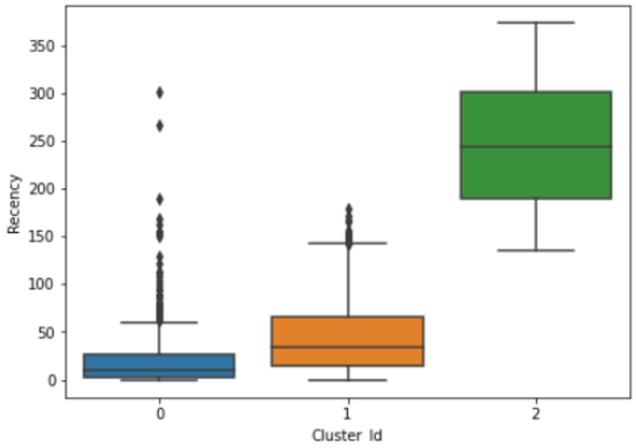


Fig 13. Box plot representing recency among different clusters

From the above graph (fig 13) it is evident that, the third cluster of customer shows higher recency in their purchase, i.e., the period among their subsequent purchases is found to be more compared to the other groups.

The cluster 0 is the most potential buyers of gifting industry, so investigating more on their choices and preferences and designing promotions and offers targeting this group would help in increasing the profit margin.

Using association pattern mining, we identified 280 associations with a minimum support of 0.02.

X. *FUTURE STUDY*

As a future extension of this project, we are planning to analyse the cluster 0 more deeply and apply the association pattern mining on this group of customers alone. We are expecting to find useful associations among the products which would talk more about the cultural lineage of this customer class and these inferences can be put together to make better marketing strategies and production strategies focusing more on this audience.

XI. CONCLUSION

The gift industry is an ever-growing market, proper market studies and forecasts and pattern mining would help in earning more profit. Identifying the target audience correctly and finding their interests and preferences is important in a successful running of a business setting. Clustering and associated pattern mining would help in this process.

XII. REFERENCES

[1]. IBM Cloud Education. (2020, August 25). Exploratory Data Analysis. Retrieved from IBM.com: https://www.ibm.com/cloud/learn/exploratory-data-analysi

[2]. Manish Kumar. (2019). Online Retail K-Means & Hierarchical Clustering. Retrieved from Kaggle.com. https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering/data